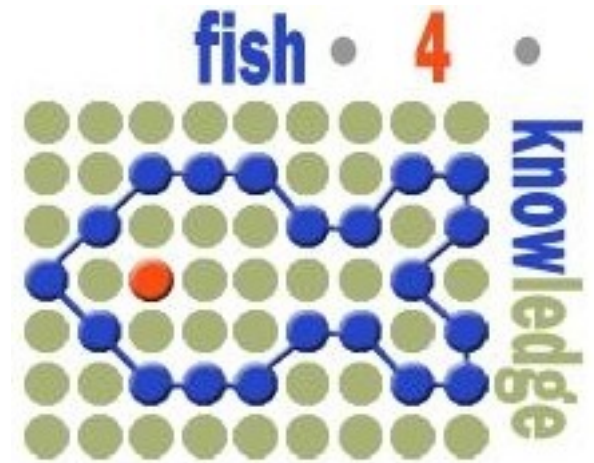


How to Efficiently Process Large Videos? A Workflow Perspective

Gaya Nadarajan
gaya.n@ed.ac.uk

Fish4Knowledge (F4K)

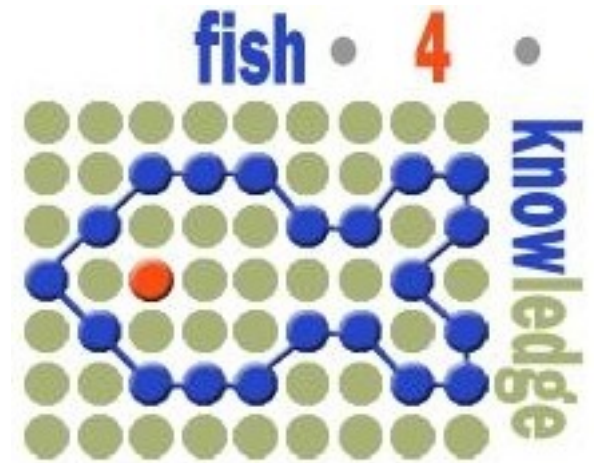
<http://www.Fish4Knowledge.org/>



- Three-year E.C. funded project, started in Oct 2010
- Partners:
 - U. Edinburgh (Image Processing – IPAB)
 - U. Edinburgh (Workflow – CISA)
 - CWI, the Netherlands (User Interface, Querying)
 - U. Catania, Italy (Image Processing)
- NCHC, Taiwan (Data Provider & HPC)

Fish4Knowledge

<http://www.Fish4Knowledge.org/>



Aims:

- Automatic storage and extraction of information of observed marine life from database
- Provide interfaces for non-expert users (marine scientists) to query on

F4K Facts

- >3000 species to be identified
- Continuous collection: 24-25K videos & counting, 8fps, 20-50MB each, 20GB per hour, 100TB per year
- 10 static cameras, resolution 680x480/1280x720 (CCTV/HDV)



video2

2007-06-01 10:00:00

20 Questions

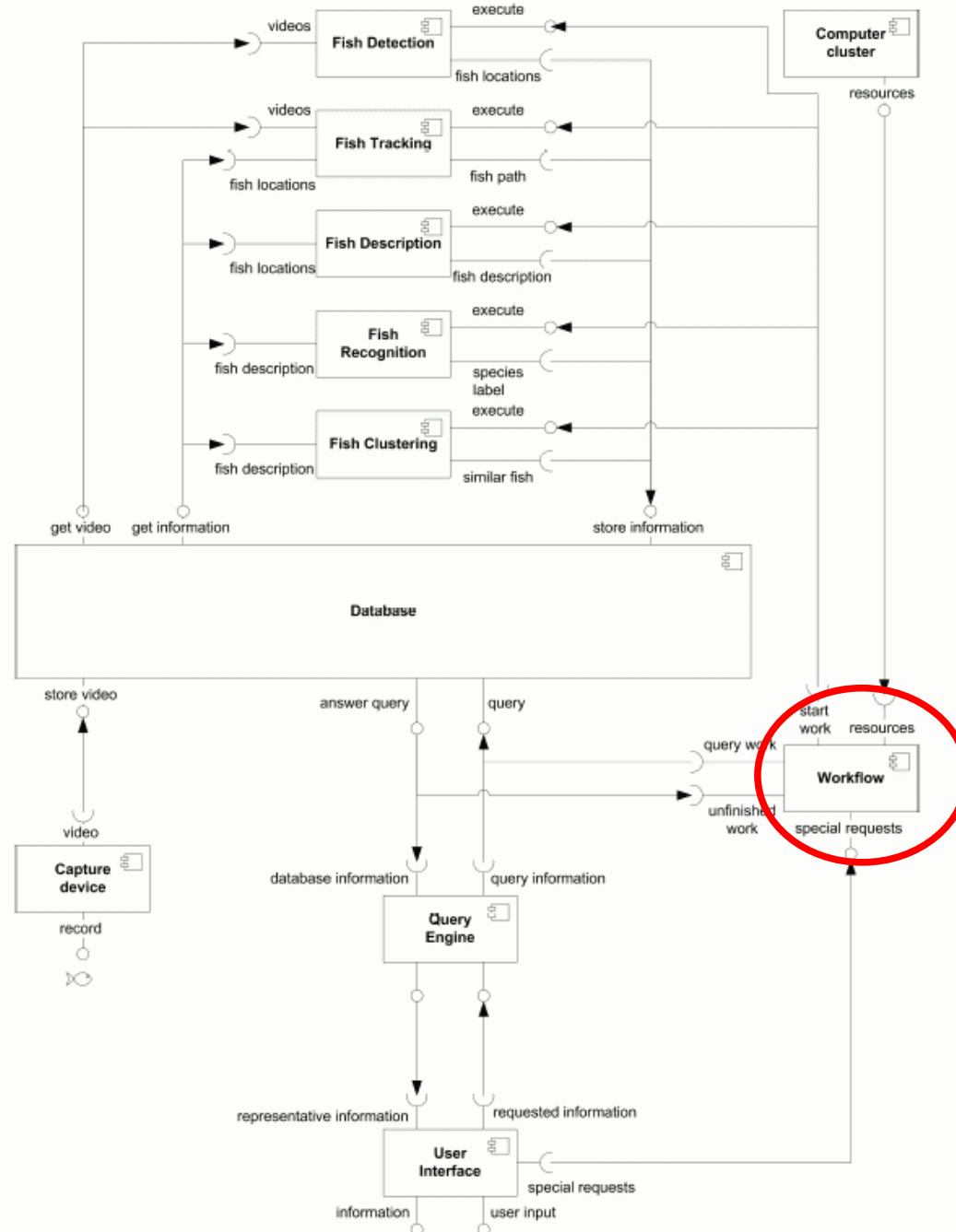
1. How many species appears and their abundance and body size in day and night including sunrise and sunset period.
2. How many species appears and their abundance and body size in certain period of time (day, week, month, season or year).
Species composition change within one period.
3. Give the rank of above species, i.e., listed according to their abundance or dominance. How many percent are dominant (abundant), common, occasional and rare species.
4. Fish color pattern change and their behavior in the night for diurnal fish and vice versa for nocturnal fishes.
5. Fish activity with one day (24 hours).
6. Feeding, predator-prey, territorial, reproduction (mating, spawning or nursing) or other social or interaction behavior of various species.
7. Growth rate of certain species for a certain colony or group of fishes.

20 Questions

8. Population size change for certain species with one period of time.
9. The relationship of above population size change or species composition change with environmental factors, such as turbidity, current velocity, water temperature, salinity, typhoon, surge or wave, pollution or other human impact or disturbance.
10. Immigration or emigration rate of one group of fishes inside one monitoring station or one coral head.
11. Solitary, pairing or schooling behavior of fishes.
12. Settle down time or recruitment season, body size and abundance for various fishes.
13. In certain area or geographical region, how many species could be identified or recognized easily and how many species are difficult. The most important diagnostic character to distinguish some similar or sibling species.

20 Questions

14. Association among different fish species or fish-invertebrates.
15. Short term, mid-term or long term fish assemblage fluctuation at one monitoring station or comparison between experimental and control (MPA) station.
16. Comparison of the different study result between using diving observation or underwater real time video monitoring techniques. Or the advantage and disadvantage of using this new technique.
17. The difference of using different camera lens and their angle width.
18. Is it possible to do the same monitoring in the evening time.
19. How to clean the lens and solve the biofouling problem.
20. Hardware and information technique problem and the possible improvement based on current technology development and how much cost they are.



Workflow Component

Roles:

- Deal with special queries (not possible by querying DB)
- Manage and schedule jobs/processes to cluster optimally
- Deal with live data streams

Considerations

- Distributed and/or parallel processing
- Batch processing
- Data streaming instead of job-scheduling to minimise volume of intermediate data

Workflow: Special Queries

- Queries need to be formalised into machine readable VIP goals and VIP solutions
- Done via ontology and planning based workflow composition mechanism (see PhD Thesis 2010)
- Assumes that multiple executables exist for VIP tasks
- Can deal with user preferences (e.g. processing time, accuracy of result) and can read in descriptions of video (e.g. brightness, clearness levels) to help with context-sensitive selection of optimal VIP modules

Optimal Job Scheduling

- Facilities: "VMware On-Demand" platform, 48 CPUs, 144G RAM
- Distribution and parallelisation considered
- Is it worth parallelising VIP tasks, if so, which ones?

Module (single CPU, 10-min video)	Ideal	Expected	Worst-case
Fish detection	real-time	20 mins	60 mins
Fish description	1s per fish	5s per fish	10s per fish
Fish classification	0.1s per fish	1s per fish	2s per fish
Fish clustering	0.1s per fish	1s per fish	2s per fish

- Both **data-** and **compute-intensive**

Options

- (1) Run your code in the pure and native VM Server : shared memory architecture (a pure 48 CPUs SMP system : x86_64)
- (2) Run your code in one VM system : shared memory architecture (a virtual SMP system : x86_64)
- (3) Run your code in several VM systems : individual and distributed memory architecture (a virtual Cluster system) , it's similar to the real PC Cluster. (MPI / PVM)

Data Streaming vs. Job-scheduling

- Is it worth looking into data streaming?
- Minimises intermediate data which is useful for VIP domain
- Job-scheduling also has ways of reducing data movement, *e.g.* clustering
- Job scheduling and then data streaming?

Adaptation for Dynamic Data

- Live streams pose changes in the data that come in within a single workflow execution
- How to model an adaptive workflow solution that can manage such scenario
- Consider Cloud solutions, *e.g.* Hadoop for dynamic scalability?